DOCUMENT RESUME

ED 126 107 TM 005 362

AUTHOR	Smith, Donald M.
TITLE	The KR-20 Reliability Coefficient as a Special Case of a More General Formula.
PUB DATE	[Apr 76]
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Comparative Analysis; Grade Point Average; Predictive Ability (Testing); Predictive Validity; Response Style (Tests); Scoring; *Statistical Analysis; Test Interpretation; *Test Reliability; *Timed Tests; *True Scores
IDENTIFIERS	*Kuder Richardson Formula 20; Speeded Tests; Test Theory

ABSTRACT

The Kuder Richardson-20 Formula is shown to be a special case, where each examinee is given sufficient time to answer each item, of a more general formula where each examinee may not be allowed the necessary time. The formula is extended to allow two scores, knowledge and speed, to be extracted from each examinees test score. Using a sample of 82 first quarter freshmen it was found that, compared to the simple total score, the two extracted scores gave better prediction of grade-point-average (gpa) in quantitative areas and was equally effective in predicting gpa in nonquantitative areas. (Author/DEP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED126107

THE KR-20 RELIABILITY COEFFICIENT AS A SPECIAL CASE OF A MORE GENERAL FORMULA

Donald M. Smith

Ball State University

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 362

A paper presented at the annual convention of The American Educational Research Association, held in San Francisco, California, April 19th - April 23rd, 1976.

Introduction

One of the basic assumptions of classical test theory is that each subject has had ample time to attempt every item in a test. Under this assumption an omitted item represents a true lack of knowledge and cannot be attributed to a failure to reach the item. It is obvious that under a restrictive time limit, defined as a period of time such that every testee does not have sufficient time to attempt every test item, this assumption cannot be met. Several persons (Gulliksen, 1951; Cronbach & Warrington, 1951; Helmstadter & Ortmeyer, 1953) have devised formulae that allow estimates to be made of the "speededness" of a test.

With the exception of one formula proposed by Cronbach & Warrington, all of these formulae are based on a single administration of one test and involve the comparison of a measure of the variance of omitted items to either the total test variance or the total "error" (usually defined as the number of omitted and incorrectly answered items) variance. Although the various approaches to the calculation of the value may vary somewhat, the basic rational underlying the formulae appears to be the same. Empirical support for the fact that they provide similar results is reported by Helmstadter & Ortmeyer in their previously cited article. Tau, an index proposed by Cronbach & Warrington, requires that parallel form of the test be administered under timed and untimed conditions. The correlations between the four obtained scores are then used to determine the proportion of the observed test variance that can be attributed to "speed."

. . . . . .

Insert formula 1 about here

. . . . . .

Although the theoretical limits of this index would approach negative and positive infinity (assuming that the correlation between the parallel forms given under the same condition were zero), the practical limits are likely to be between -3.00 (in those cases where both r_{12} and r_{34} are both 1.00 and r_{14} and r_{23} are both 0.50) and zero (when all four correlations are the same value). This ability to have negative values is a desirable characteristic that is not possessed by any of the other proposed indices. All of the other formulae provide estimates of the proportion of test variance that may be attributable to "speed" that have a lower limit of zero; and, although this is never specifically stated, assume that the sole effect of a restrictive time limit is to increase the total test variance: and hence, by definition, the estimated reliability of the test. While this is certainly a likely outcome it does not automatically follow that it will always be so.

The actual effect of a restrictive time limit on the test statistics will vary, depending upon the statistic being considered, the degree of speededness, and the characteristics of the group taking the test. Nunnally (1967, p.566) provides a good summary of these effects.

> "The potential effects of restrictive time limits on the mean score are obvious. If there are any effects at all, the expectation is that the mean will increase with increasing fractions of

> the comfortable-time, with little increase being expected above 100 percent of the comfortable-time. There is, however, no strict relationship between the mean and reliability or validity. A mean near the center of the usable score range tends to favor high reliability, but the relationship holds in only a loose statistical way."

And Morrison points out the problems associated with indices that are based on a single administration of one form.

> "The major difficulty faced by all single-trial indices is that any score which might be used (right, wrong, number attempted, number right divided by the number attempted) is psychologically complex, in that it may reflect both speed and ability influences under time-limit administration. We simply cannot tell from time-limit data alone what the effects of the time-limit will be."

Table 1 illustrates what may occur to certain test statistics when a restrictive time limit is employed. To construct the table two testee characteristics were considered. The first of these was ability, defined as the proportion of correct responses that a person could correctly answer under untimed conditions. Two ability levels, 0.9 and 0.4, were used. The second characteristic was the speed at which a testee could work and this was defined as the proportion of questions that he could answer in the allowed time. There were also two levels, 0.9 and 0.4, of this characteristic. Twenty-three different distributions of testees were considered to have taken a 100 item test. The table gives these distributions and the expected values of three test statistics under both timed and untimed administrations. It should be noted that the untimed administrations, give results that may be considered as representing the "true" state of affairs.

. . . . . .

Insert Table 2 About Here

. . . . . .

In all cases the obtained mean is a low estimate of the value that would be obtained under pure power conditions; and, as the correlation of 0.683 between the means obtained under timed and untimed conditions shows, the relationship is not such as would allow much confidence to be placed on the values obtained under timed conditions. Much the same is true, albeit to a far greater extent, with the estimates of total test variance. The relationship between the two sets of variances is only 0.189 and the timed values may be either more or less than the untimed values. It is this fact that makes Cronbach's proposed index, tau, so attractive. For it is the only index that allows for the possibility that the total test variance may be reduced when the test is given with less than a comfortable time limit.

Purpose of a Test

Before discussing the assumptions that underly the use of the KR20, and the modification of the formula that might make it possible to minimize the consequences of the violation of the assumptions, it might be useful to briefly review the purposes that a test is to serve. Although there may be disagreement on this point it is the writers belief that a test score serves but one purpose: to estimate the testees present position on a behavioral continuum. The purposes why it is desired to make this prediction may vary, but will usually involve some type of comparison. There will always require the comparison of an entities present position with either the previous position of the same, or the previous position of another, entity. An entity may be either an individual or a group; in the latter case the mean value of the group would be used.

The purpose of the reliability coefficient is thus quite clear. It allows the person making the comparison(s) to make a judgement as to how well the positions on the continuum have been determined: a judgement of particular importance whenever it becomes necessary to interpret the findings of a study whose goals were to determine whether there were differences between two or more estimates of position. A state of affairs such as would exist under the first condition given in Table 1, and presented in Table 2, would be highly undesirable: since, under these conditions, only 25 percent of the subjects would have their actual true score included within the 95 percent confidence limits about their estimated true score. One can only wonder how many negative findings, or failure to replicate previous results, may have been caused by conditions similar to this.

. . . . . .

Insert Table 3 About Here

. . . . . .

The Assumption Underlying the Interpretation of the KR20 Reliability Coefficient

The KR20 reliability coefficient can be calculated by any of several different formulae: two of which are given below.

. . . . . .

Insert Formulae 2 & 3 About Here

. . . . . .

Simply stated, and without going into a detailed psychometric explanation, the KR20 reliability coefficient can be considered as the correlation between a persons observed test score and his "true" score on the domain that the test purports to measure. Formula 2 expresses this in terms of the inter-correlations of the k items that make up the test; while formula 3, which is mathematically equivalent to 2, uses the total test variance and the sum of the item variances.

Since the formulae are based on population parameters the assumption that must be met whenever sample estimates are used is that these sample statistics must be unbiased estimates of their corresponding population parameters. It would appear that, in the case of the KR20, two conditions must be satisfied if the assumption is to be met. The first of these is relatively easy to satisfy and requires only that the subjects be a random sample from the population of interest. The second condition will be satisfied whenever each person in the sample has been given sufficient time to allow him to answer every item to which he knows the answer. To the extent that this is not so the sample statistics will be biased estimates of their associated parameters, with the precise nature of the bias being completely unknown. In cases such as this not only would the calculated value of the KR20 be uninterpretable, but the test scores of those subjects who had insufficient time to complete the test would be unknown. Table 4, which will also be used to illustrate a modification of the traditional KR20 formula that would adjust for the effects of a restrictive time limit, demonstrates the indeterminacy of certain of the required test statistics under a speeded test administration.

. . . . . .

Insert Table 4 About Here

. . . . . .

There are, in all cases, n plus k pieces of information required for the calculation of the necessary statistics: the k column sums that will provide estimates of the item variances, and the n row sums that allow the estimated test variance to be calculated. In the example 17 of the 30 pieces of information (four of the row totals and thirteen of the column totals) are indeterminate. The usual practice is to mark an omitted item as wrong: assuming that the omission was caused by a lack of knowledge as to the correct answer and not by a possible failure to teach the item. The item difficulty is thus defined as the number of correct responses divided by the number of subjects. This is equivalent to assuming that the speed with which a person can answer test questions is perfectly correlated with his knowledge of the material to which the test questions pertain. This assumption is, in no way, supported by a rather extensive body of research (Smith, 1971).

It would, however, be possible to estimate the item variances by using only the available information. The estimated item difficulty would then be the number of correct responses divided by the number of persons who reached the item. This would certainly be a sound practice in that the obtained values would be based only on available information and are unbiased estimates (although some of them, being based on rather small samples, might have quite wide confidence limits) of the population values. The same procedure cannot, unfortunately, often be used to obtain estimates of the raw sums: the persons total test score. This is caused by the widespread practice among American test constructors of

arranging the items within a test in order of increasing difficulty (i.e., the easiest item is placed first, followed by the next easiest, and so on to the most difficult item). There is thus a confounding of item difficulty with item placement, and the estimation procedures used with the item variances would require an assumption that is, by the definition of the procedure used to determine item placement, impossible: the probability of correctly answering an item is independent of item placement.

There are two approaches that may be used to solve the problem. The first of these makes use of the item inter-correlations formula and uses only the available data to estimate the correlations. Any correlation program that will handle missing values could be used to carry out the necessary calculations. The estimated reliability of the test could then easily be calculated. The unweighted sum of the item means would serve as the best estimate of the mean score of the population from which the sample was drawn. This approach appears to be quite useful for those cases where it is desired to estimate the position of a group on the behavioral continuum of interest. It would not, however, be of any value for those instances where it was desired to make statements about the position of individuals.

The second approach would require unbiased estimates of the probability that a person would have correctly answered unreached items had he been given the opportunity to do so. This is, in essence, a sampling problem; and either of two methods may be used to obtain the required estimates. Both methods require that the traditional American procedure of arranging items in order of increasing difficulty (i.e., the easiest item first, followed by the next easiest item, and so on) could no longer be used. In the opinion of the author this would cause no great harm, since there does not appear to be any meaningful reason for this type of item arrangement. The first of these methods, which is the simplest, requires that the items be randomly assigned to their position within the test. This would allow an estimate to be made of a subjects test score by using formula 4: which defines that score as the number of correct responses, divided by the

.....

Insert Formula 4 About Here

.....

number of items attempted, and multiplied by the number of items in the test. This method of estimation is hereafter referred to as condition 2 and its use is illustrated in Table 3.

.....

Insert Table 3 About Here

.....

The other approach to item placement, hereafter called condition 3, is somewhat more complex in that it employs stratified, rather than simple, random

item placement. The procedure is commonly used in England and appears to have been first proposed by Ellis in 1928. The advantages associated with its use are the same as when any stratified random sample is used: greater precision of the resulting estimates. Under this procedure the items are first categorized into m levels of difficulty. One item from each difficulty level is then randomly selected and these m test items, arranged in ascending order of difficulty, are the first m items in the test. The procedure is repeated, in turn, until all items have been placed in the test. The test will thus consist of a series of cycles each of which contains m test items. Each cycle is therefore a representative sample of the entire range of tasks within the domain being measured by the test. To illustrate how this would work consider a 100 item test in which each item has been assigned to one of five levels of difficulty: 0.99-0.81, 0.80-0.61, 0.60-0.41, 0.40-0.21, and 0.20-0.01. There would thus be, in this case, twenty complete cycles and it would be a relatively simple matter to estimate a subjects performance on any unreached items. The person by items response matrix that is presented in Table 3 consists of four cycles of five items.

Using this method a subjects test score can be viewed as consisting of m separate parts: where m is the number of item difficulty levels (strata) used in the test. His score, in those cases where all items were not reached, is thus the sum of the m adjusted level scores. Each of these level scores (formula 5), is calculated by dividing the number of correct responses to the items within a level by the number of items within that cycle that were reached, and multiplying the result by the total number of items within that cycle. Two points should be mentioned at this time. First, it is not necessary that each strata contain the same number of items, although the computations required become easier if this is the case; and, second, both of the methods described above provide results that are identical with those obtained from the traditional method of computing test scores, whenever all subjects have been given sufficient time to attempt each test item. It follows that the various test statistics will also be the same under such conditions and the traditional formula for calculating the KR20 can thus be viewed as a special case for a more general formula that estimates the responses to unreached items by using the available information of that subject.

As the example that is presented in Table 3 illustrates, the results obtained by the two modified formulae are in very close agreement: and both differ considerably from those obtained when the omitted responses are treated as wrong answers. In the example the modified formulae give increased values for the total test variance, the sum of the item variances, and the reliability of the test. The standard error of measurement is thus, for the modified formulae, smaller. The correlations between the three sets of scores, which are also given in Table 3, are quite interesting.

In order to determine what might happen to the various test statistics in other circumstances an analysis of forty tests that had been submitted to the Office of Examination Services was carried out. The various test statistics were computed for each test by two different methods: treating omits as incorrect responses (condition 1) and treating omits as unreached items under the assumption

of random item placement (condition 2). The results of this analysis are given in Table 4 and support the earlier statement that it is impossible to say, in advance and for any given test, what will happen to the test statistics.

.....

Insert Table 4 About Here

.....

Predictive Validity of the Modified Estimates

One final analysis was carried out to determine if there were any differences in the ability of two test scores, one obtained from the traditional method and one obtained from the modified method (stratified item placement), to predict various criteria. The subjects were the 82 first term freshman enrolled in a required course in introductory psychology at a large mid-western state university. Each subject was administered the Verbal and Quantitative sub-tests of the College Qualification Test. This is a commercial test published by The Psychological Corporation. The Verbal sub-test consists of 75, four choice, verbal analogies while the Quantitative sub-test consists of 50, four choice, mathematical questions. Although the tests are stated to be untimed there is a recommended time limit and this was used. Both of the tests had, on the basis of item difficulty data previously supplied by the publisher, been modified into condition 3 (stratified item placement). Each cycle contained five test items. Each subject thus had six different scores: the CQT-V and CQT-Q computed using the traditional (condition 1) formula; the CQT-V and CQT-Q computed using the stratified (condition 3) formula; and two estimates of subject speed. These last two scores were calculated by dividing the number of items that had been reached by the number of items in the test: and the possible values range from 0.00 to 1.00. These were used as crude estimates of the rate at which a subject could perform the tasks sampled by the test items and may be considered, in a loose sense, as the speed with which a person can handle new information. These six scores, along with the sex of each subject, were then used as predictor variables for three different academic criteria: first term grade point average in non-mathematics/science courses (GPA1), first term grade point average in mathematical/science course (GPA2), and first term grade point average in all courses (GPAT). The results of this series of analyses are given in Table 5.

.....

Insert Table 5 About Here

.....

Sex was not a significant predictor for either of the three criteria. As would be expected, the verbal test score were the best single predictors of GPA1 and the quantitative tests scores were the best single predictors of GPA2. The multiple correlations were, in all cases, larger when the modified scores were used as predictors. It was quite interesting to note that both of the speed indices were significant predictors of GPA2. This could be interpreted as meaning that, for the type of material covered in these courses, the rate of

response (taken as an indication of the rate of learning acquisition) is an important factor. The two sets of test statistics are quite similar. This is to be expected if, as is the case with the CQT, the time allowed is close to being sufficient for all subjects to attempt all items. This analysis gives tentative support to the contention that more accurate, and therefore more useful, information is provided by the modified formulae.

Recommendations

Based upon the series of analyses herein reported it would appear that at least six recommendations are in order.

1) All commercial test publishers should routinely provide information as to the degree of speededness associated with their tests. In the case of multi-level tests (i.e., those tests that are used for several different age groups), especially those for use in the earlier grades, the information should be provided for each level at which the test may be used. Should there be reason to believe that sub-divisions of the population differ in their response rates then this information should also be provided for the sub-divisions.

2) Research as to which of the various indices is most accurate should be carried out. In the interim any of the single administration indices referenced in this paper could be used to provide the needed information.

3) Firms providing test analysis and reporting services should routinely calculate, and provide as part of their services, the speed index of each group administration of a test. This is a trivial problem of computer programming and requires a sub-routine of less than twenty statements.

4) Careful consideration should be given as to whether the present procedure used to determine item location within a test should be changed. There is very little practical or theoretical reason to retain the present procedure; although there are several benefits that would follow the adoption of either a simple random, or a cyclical, arrangement of items.

5) Should the above be adopted the firms mentioned in 3) should also provide, if requested, test statistics, including the individual test scores based on the appropriate modified formula. Although the programming required to provide this service is less trivial than was the case with the speed index, it is still a very simple matter.

6) A program of research aimed at discovering whether the rate of response is indeed a separate measurable dimension should be initiated. Should this prove to be the case, and should response rate be related to the rate of intellectual development, and the existing research in this area indicates that this may well be the case, the implications for education are self evident.

Summary

Several indices that have been proposed as estimates of the degree of speededness of a test were discussed. With one exception, Cronbachs tau, all of these appear to be based on an unsupportable rational: that the effects of a restrictive time limit will be to, in all cases, increase the total test variance. The effects of the speeded administration of a test were shown to result in test results that are basically uninterpretable: with the problem being caused by the bias that is introduced into the various test statistics as a result of the insufficient time limit. It was further demonstrated that the traditional KR20 formula is a special case, requiring the assumption that all subjects have been allowed sufficient time to attempt all of the test items, of a more general formula that does not require this restrictive assumption.

An empirical study indicates that the scores provided by the modified formula were slightly better predictors of first term grade point average than were those scores provided by the traditional formula.

Six recommendations, especially applicable to firms publishing tests or providing test analysis services, were given.

REFERENCES

Cronback, L. J. & Warrington, W. G. Time-limit tests: Estimating their reliability and degree of speeding. Psychometrika, 1951, 16.

Cronback, L. J. & Warrington, W. G. Time-limit tests: Estimating their reliability and degree of speeding. Psychometrika, 1951, 16, 167-188.

Eliis, R. S. A method of constructing and scoring tests with time limits to eliminate or weigh the effect of speed. School and Society, 1928, 28, 205-207.

Gulliksen, H. The reliability of speeded tests. Psychometrika, 1950, 15, 259-269.

Helmstadter, G. C. & Ortmeyer, D. H. Some techniques for determining the relative magnitude of speed and power components of a test. Educational and Psychological Measurement, 1953, 13, 280-287.

Mollenkopf, W. G. Time limits and the behavior of test takers. Educational and Psychological Measurement, 1960, 20, 223-230.

Morrison, E. J. On test variance and the dimensions of the test taking situation. Educational and Psychological Measurement, 1960, 20, 231-250.

Nunnally, J. C. Psychometric Theory. McGraw-Hill, New York, 1967.

Smith, D. M. The validity of factor score estimates of speed and accuracy as predictors of first term grade point average. (Doctoral dissertation, Florida State University) Ann Arbor, Michigan: University Microfilms, 1971.

Toops, H. A. A comparison, by work-limit and time-limit, of item-analysis indices for practical test construction. Educational and Psychological Measurement, 1960, 20, 251-266.

Wesman, A. G. Some effects of speed in test use. Educational and Psychological Measurement, 1960, 20, 267-274.

Table 1
Expected Values of Test Statistics
Under Timed and Untimed Conditions for
Groups of Differing Characteristics

Proportion of Sample				Timed Administration			Untimed Administration			
A=.9		A=.4								
S=.9	S=.4	S=.9	S=.4	Mean	Variance	KR21	Mean	Variance	KR21	No
.0.25	0.25	0.25	0.25	42.25	567.19	.967	65.00	625.00	.973	1
.0.50	0.00	0.50	0.00	58.50	506.25	.962	65.00	625.00	.973	2
0.00	0.50	0.00	0.50	26.00	100.00	.816	65.00	625.00	.973	3
0.50	0.50	0.00	0.00	58.50	506.25	.962	90.00	0.00	I	4
0.00	0.00	0.50	0.00	26.00	100.00	.816	40.00	0.00	I	5
.0.50	0.00	0.00	0.50	48.50	1056.25	.986	65.00	625.00	.973	6
0.00	0.50	0.50	0.00	36.00	0.00	I	65.00	625.00	.973	7
1.00	0.00	0.00	0.00	81.00	0.00	I	90.00	0.00	I	8
0.00	0.00	1.00	0.00	36.00	0.00	I	40.00	0.00	I	9
0.00	1.00	0.00	0.00	36.00	0.00	I	90.00	0.00	I	10
0.00	0.00	0.00	1.00	26.00	0.00	I	40.00	0.00	I	11
0.70	0.00	0.30	0.00	67.50	425.25	.958	75.00	525.00	.974	12
.0.30	0.00	0.70	0.00	49.50	425.25	.951	55.00	525.00	.962	13
0.70	0.30	0.00	0.00	67.50	425.25	.958	90.00	0.00	I	14
0.30	0.70	0.00	0.00	49.50	425.25	.951	90.00	0.00	I	15
0.00	0.00	0.70	0.30	30.00	84.00	.758	40.00	0.00	I	16
0.00	0.00	0.30	0.70	22.00	84.00	.812	40.00	0.00	I	17
0.00	0.30	0.00	0.70	22.00	84.00	.812	55.00	525.00	.962	18
0.00	0.70	0.00	0.30	30.00	84.00	.758	75.00	525.00	.974	19
0.70	0.00	0.00	0.30	61.50	851.41	.982	75.00	525.00	.974	20
0.30	0.00	0.00	0.70	35.50	887.25	.984	55.00	525.00	.962	21
0.00	0.30	0.70	0.00	36.00	0.00	I	55.00	525.00	.962	22
0.00	0.70	0.30	0.00	36.00	0.00	I	75.00	525.00	.974	23

$r_{mt,mu} = 0.6825$ $r_{vt,vu} = 0.1890$

Table 2
True Scores, Estimated True Scores and Confidence Limits for Sample Number 1

Sub-Sample Characteristics	Observed Score	Estimated True Score	95 Percent Confidence Limits	True Score
A=.9, S=.9	81	79.72	71.24 - 88.20	90
A=.9, S=.4	36	36.21	27.73 - 44.69	90
A=.4, S=.9	36	36.21	27.73 - 44.69	40
A=.4, S=.4	16	16.87	8.39 - 25.35	40

Table 3 (Concl'd)

Scoring Condition	Mean	Standard Deviation	Sum of the Item Variances	KR-20 Reliability	Standard Error of Measurement
Traditional (Cond 1)	10.10	3.477	4.31	.715	1.856
First Modified (Cond 2)	12.89	4.410	4.50	.854	1.685
Second Modofied	12.93	4.409	4.50	.854	1.685

Inter-correlations Among the Three Sets of Scores

	C1	C2	C3
C1	1.000	0.552	0.552
C2	0.443	1.000	1.000
C3	0.477	0.999	1.000

Note: product moment above the diagonal
rank order below the diagonal

Table 4

Comparison of Test Statistics Obtained From the Analysis of Forty Tests Using the Traditional KR20 Formula and the Simple Random Item Placement Modification

Test Variance	Sum of the Item Variance	KR20 Reliability	Number of Occurences
Increase	Increase	Increase	3
Increase	Increase	Decrease	5
Increase	Decrease	Increase	6
Increase	Decrease	Decrease	0
Decrease	Increase	Increase	0
Decrease	Increase	Decrease	15
Decrease	Decrease	Increase	6
Decrease	Decrease	Decrease	5

Note: The changes given are those of the modified formula, referenced to the traditional formula

Table 3

Illustrative Person by Items Response Matrix
With Item and Test Statistics and Individual Test Scores
Computed Using Each of the Three Methods

		Items																				Individual Responses				Individual Test Scores		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	R	W	O	A	C1	C2	C3
	a	1	1	1	0	1	1	-	1	1	1	0	1	1	1	0	1	o	o	o	o	12	4	4	16	12	15.0	15.0
	b	1	0	0	1	1	1	0	1	1	0	0	1	0	1	1	0	1	0	0	1	11	9	0	20	11	11.0	11.0
	c	1	1	0	1	1	0	1	0	1	1	1	1	0	1	0	1	o	o	o	o	10	5	5	15	10	13.3	13.3
	d	0	1	1	1	1	1	1	1	0	1	1	o	-	o	o	o	o	o	o	o	9	2	9	11	9	16.4	16.7
Persons	e	1	0	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	6	14	0	20	6	6.0	6.0
	f	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	18	2	0	20	18	18.0	18.0
	g	0	1	1	0	1	-	0	0	1	-	-	0	0	0	1	1	0	0	o	o	6	12	2	18	6	6.7	7.0
	h	1	0	1	1	0	0	1	0	1	0	0	1	0	0	1	1	0	1	0	o	9	10	1	19	9	9.5	9.3
	i	1	1	1	1	1	1	1	o	o	o	o	o	o	o	o	o	o	o	o	o	7	0	13	7	7	20.0	20.0
	j	1	0	0	1	1	1	0	1	0	1	1	0	1	0	1	1	1	0	1	1	13	7	0	20	13	13.0	13.0
	R	8	6	7	7	8	6	6	6	6	5	5	4	4	3	6	5	3	2	2	2							
Item	W	2	4	3	3	2	4	4	4	3	3	4	4	4	4	5	2	3	4	3	2							
Statistics	O	0	0	0	0	0	0	0	1	1	1	1	2	2	2	2	3	4	4	5	6							
(decimals	A	10	10	10	10	10	10	10	9	9	9	9	8	8	8	8	7	6	6	5	4							
omitted in	P1	80	60	70	70	80	60	60	60	60	50	50	40	40	30	60	50	30	20	20	20							
last four	V1	16	24	21	21	16	24	24	24	24	25	25	24	24	21	24	25	21	16	16	16							
rows)	P2	80	60	70	70	80	60	60	67	67	56	56	50	50	38	75	71	50	33	40	50							
	V2	16	24	21	21	16	24	24	22	22	25	25	25	25	24	19	21	25	22	24	25							

Key:

1: a correctly answered item
0: an incorrectly answered item
-: an omitted item that was assumed to be incorrect
o: an omited item that was assumed to not have been reached
A: the number of attempted items (individual) or the number of persons attempting an item
C1: the test scoreof an individual under traditional scoring rules (condition 1)
C2: the test score of an individual under the first modified scoring rule (condition 2)
C3: the test score of an individual under the second modified scorinf rule (condition 3)
O: the number of omitted items (individual) or the number of times that an item was omitted
P1: the proportion of persons correctly answering an item under traditional scoreing rules
P2: the proportion of persons correctly answering an item under the modified scoring rules
R: the number of items correctly answered (individual) or the number of correct responses to an item
V1: the variance of an item under the traditional scoring rules
V2: the variance of an item under the modified scoring rules
W: the number of incorrectly answered items (individual, or the number of times an item was incorrectly answered

Table 5

Results of the Predictive Study
Including Variable Means and Standard Deviation
and Test Statistics

Statistic	Variable						
	CQTV	CQTQ	MCQTV	MCQTQ	GPA1	GPA2	GPAT
Mean	35.220	22.549	48.901	24.353	2.620	2.589	2.607
Stnd Dev	13.058	7.868	13.510	7.835	0.669	0.734	0.620
Sum item var	17.791	11.254	16.381	11.697			
KR-20	0.908	0.835	0.923	0.826			
SE Meas	3.966	3.197	3.760	3.265			

Variable Inter-correlations

	Verbal GPA	Math GPA	Total GPA	CQTV	MCQTV	Verbal Speed	CQTQ	MCQTQ	Math Speed
Verbal GPA	1.00	0.63	0.65	0.59	0.57	0.45	0.29	0.33	0.27
Math GPA	0.63	1.00	0.69	0.40	0.31	0.30	0.52	0.61	0.30
Total GPA	0.65	0.69	1.00	0.58	0.57	0.33	0.28	0.33	0.27
CQTV	0.59	0.40	0.58	1.00	0.85	0.50	0.34	0.41	0.26
MCQV	0.57	0.31	0.57	0.85	1.00	0.56	0.19	0.23	0.21
Verb Speed	0.45	0.30	0.33	0.50	0.56	1.00	0.17	0.21	0.20
CQTQ	0.29	0.52	0.28	0.34	0.19	0.17	1.00	0.92	0.70
MCQTQ	0.33	0.61	0.33	0.31	0.23	0.21	0.92	1.00	0.71
Math Speed	0.27	0.30	0.27	0.26	0.21	0.20	0.70	0.71	

Regression Coefficients

	Criterion					
	Traditional (Cond 1) Scores			Modified (Cond 3) Scores		
Predictor	Verbal GPA	Math GPA	Total GPA	Verbal GPA	Math GPA	Total GPA
Intercept	1.564	1.184	1.645	0.935	1.063	1.050
	(0.059)	(0.066)	(0.055)	(0.059)	(0.060)	(0.054)
CQTV	0.030	0.014	0.027			
	(0.005)	(0.006)	(0.004)			
MCQTV				0.025	-	0.024
				(0.005)	-	(0.004)
Verb Speed				-	0.513	-
				-	(0.221)	-
CQTQ	-	0.040	-			
	-	(0.009)	-			
MCQTQ				0.018	0.073	0.016
				(0.008)	(0.011)	(0.007)
Math Speed				-	-0.900	-
					(0.373)	
R	0.589	0.573	0.579	0.603	0.671	0.603
R^2	0.346	0.328	0.336	0.364	0.450	0.364

Note: Standard errors are in parentheses beneath the associated coefficient

Formulae

(1) $$\text{Tau} = 1 - \frac{r_{12}r_{34}}{r_{14}r_{23}}$$

Where: Tau is the proportion of the total test variance that can be attributed to speed

r_{12} is the correlation between Form A given under timed conditions and Form B given under untimed conditions

r_{34} is the correlation between Form A given under untimed conditions and Form B given under timed conditions

r_{14} is the correlation between Forms A & B when both are given under timed conditions

r_{23} is the correlation between Forms A & B when both are given under untimed conditions

(2) $$\text{KR20} = \frac{K}{K-1}\left[\frac{\sum\sum r_{jk} - K}{\sum\sum r_{jk}}\right] \qquad \text{and } j=1,K;\ k=1,K$$

Where: K is the number of items in the test

r_{jk} is the correlation between the j^{th} and the k^{th} test items: with the values calculated under the assumption that all omitted items are incorrect responses

(3) $$\text{KR20} = \frac{K}{K-1}\left[\frac{V(X) - \sum V(k)}{V(X)}\right] \qquad \text{and } k=1,K$$

Where: N is the number of persons who took the test

K is the number of items in the test

X is the total test score of a person on the test. This is the number of test items that were correctly answered and all omitted items are counted as incorrect responses

V(X) the variance of the N total test scores

V(k) the variance of the k^{th} test item

(4) $$XC2_n = \frac{K\sum C_{nk}}{\sum A_{nk}} \qquad \text{and } k=1,K$$

Where: $XC2_n$ is the score of the n^{th} person who took the test adjusted under the assumption of simple random item placement

N is the number of persons who took the test

n is the n^{th} person who took the test

K is the number of items in the test

k is the k^{th} item in the test

C_{nk} is the number of items correctly answered by the n^{th} person

A_{nk} is the number of items reached (attempted) by the n^{th} person

(5) $$XC3_n = \frac{\sum I_m \sum C_{nmi}}{\sum A_{nmi}}$$ and m=1,M; n=1,N; i=1,I

Where:

$XC3_n$	is the score of the n^{th} who took the test adjusted under the assumption of stratified random item placement
N	is the number of persons who took the test
n	is the n^{th} person who took the test
M	is the number of item difficulty levels (strata) in the test
m	is the m^{th} item difficulty level of the test
I_m	is the number of test items in the m^{th} item difficulty level
C_{nmi}	is the number of items in the m^{th} difficulty level that the n^{th} person correctly answered
A_{nmi}	is the number of items in the m^{th} difficulty level that the n^{th} person reached (attempted)